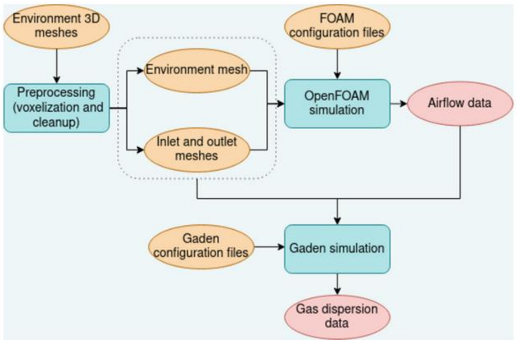
Integration of the Virtual Gas Release Dataset into the DDPM

Ayoub Ouederni

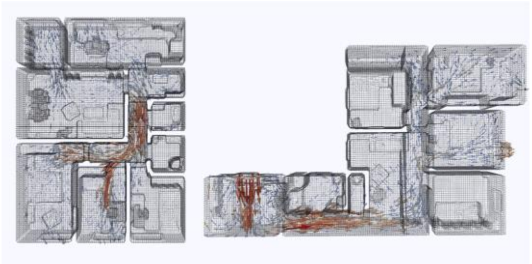
Professor : Alcherio Martinoli
Assistant(s) : Wanting Jin

Gas source localization is a key challenge in mobile robotics, with applications in environmental monitoring and industrial safety. Recent learning-based approaches have shown promising results, but their performance strongly depends on the quality and diversity of the training data. In the existing Data-Driven Plume Model (DDPM) framework used at DISAL, the learning model was trained on simplified rectangular environments with basic obstacle configurations.

The objective of this project was to extend the data capabilities of the existing pipeline by preparing gas dispersion data from more complex environments, enabling future training beyond simplified scenarios. Rather than modifying the model architecture, the focus was placed on data preparation, formatting, and compatibility, which are critical steps for scaling learning-based robotic systems.



Simulation pipeline combining airflow (OpenFOAM) and gas dispersion modeling (GADEN).



Examples of complex indoor environments from the VGR dataset, with diverse layouts and gas dispersion patterns.

To achieve this, I worked with the Virtual Gas Release (VGR) dataset, which provides gas dispersion simulations generated in realistic indoor environments using GADEN and Webots. The raw simulation outputs consist of CSV files containing gas concentration values and 3D occupancy grids. I designed a preprocessing pipeline that converts these raw files into MATLAB .mat datasets fully compatible with the DeepCFD learning framework used in the DDPM project.

The pipeline extracts physically consistent 2D slices from the 3D environments, aligns spatial representations, and formats the data to match the exact input–output structure expected by the U-Net model. In particular, the prepared inputs include obstacle maps, signed distance functions, and source-related features, while the outputs represent gas concentration fields on a fixed-resolution grid. This ensures that the new dataset can be seamlessly integrated into the existing training process without altering the model design.

Due to time constraints, retraining the model on the augmented dataset and performing quantitative evaluation were outside the scope of this project. However, this work establishes a robust and reusable data integration pipeline that enables future training on richer and more realistic environments.